

Mean Birds: Detecting Aggression and Bullying on Twitter

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn
Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali



*Web Science
Troy, New York, 2017*

WebSci'17

Social Networking Services

The Facebook logo, consisting of a solid blue rectangle with the word "facebook" in white, lowercase, sans-serif font.The ASK.fm logo, with "ask" in a bold, red, lowercase, sans-serif font and ".fm" in a smaller, red, italicized, lowercase, sans-serif font.The YouTube logo, with the word "You" in a black, sans-serif font and "Tube" in a white, sans-serif font inside a red rounded rectangle.The Yahoo! Answers logo, with "YAHOO!" in a purple, sans-serif font and "Answers" in a smaller, black, sans-serif font below it.

Social networking applications contain user profiles, variety of resources, and activities.



- A microblogging service
- Sharing of up to 140-character messages
- Sharing of any kind of content



Cyberbullying

- It is bullying that takes place using **electronic** technology
 - For the **teenagers** it is highly possible to be subject to bully behaviors
-

- **70.6%** of young people say they have seen bullying in their schools
- **9%** of students in grades 6–12 experienced cyberbullying
- **15%** of high school students (grades 9–12) were electronically bullied in the past year



Cyberbullying vs. Cyberaggression

- **Cyberaggression:** purposefully saying or doing something to hurt someone once
- **Cyberbullying:** intentionally aggressive behavior, repeated over time, that involves an imbalance of power



Crawling from June to August 2016:

- **Baseline:** 1M random tweets
- **Hate-related:** 650K tweets based on 309 bully- and hate-related hashtags

309 hashtags: **#GamerGate** and 308 co-appeared ones

Gamergate Controversy

- A coordinated campaign of harassment in the online world
- It involves sexism, feminism, and “social justice” and takes place on social media like Twitter

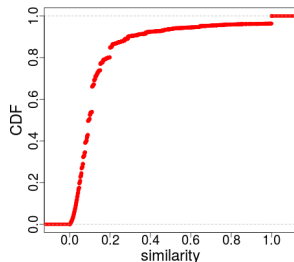
#GamerGate



Gamergate controversy provides us a unique point of view into online harassment campaigns

Preprocessing

- **Cleaning:** Removal of stop words, punctuations marks
 - **Spam removal:** Based on the number of hashtags, and duplications
-
- Avg. # hashtags: 0 to 17
 - Hashtags: we set the limit to 5
 - Similarity: Levenshtein distance
 - About 5% of the users with highly similar posts



How do you characterize the following user?

1/10

User profile description: -

Tweet 1 *your profile pic sucks! U should wear a mask to hide from the sun ☹️☹️☹️☹️*

Tweet 2 *Our class prom night just got ruined because u showed up. Who invited u anyway?*

Tweet 3 *Some1 should stalk u and have fun with u..;*

Tweet 4 *Don't cry... U can do shit about it...No matter what u do, your pics are out there.:D*

Tweet 5 *Why do you even show up at school?Nobody cares and neither should u!*

☒ Aggressive user ☐ Bullying user ☐ Spammer ☐ Normal user

NEXT

Aggressive

Someone who posts at least one tweet or retweet with negative meaning, with the intent to harm or insult other users (e.g., the original poster of a tweet, a group of users, etc.).

Bullying

Someone who posts multiple (at least two) tweets or retweets with negative meaning for the same topic and in a repeated fashion, with the intent to harm or insult other users (e.g., the original poster of a tweet, a minor, a group of users, etc.) who may not be able to easily defend themselves during the postings.

Spammer

Someone who posts tweets or retweets of advertising/marketing or other suspicious nature, such as to sell products of adult nature, phishing attempts, etc.

Ground Truth

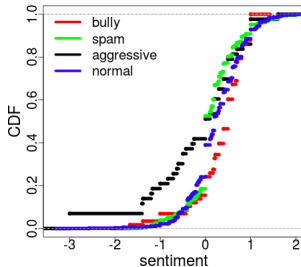
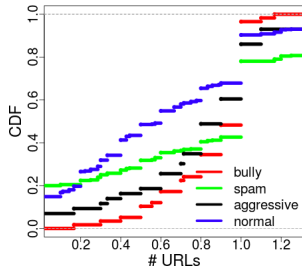
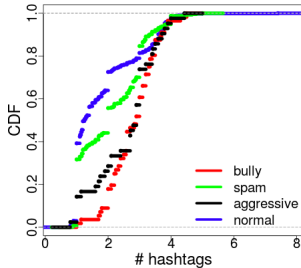
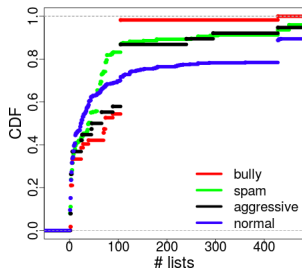
- #users: 1,307
- #tweets: 9,484
- Users' distribution in classes: 4.5% bullies, 3.4% aggressors, 31.8% spammers*, 60.3% normal

* someone who posts texts of advertising/marketing or other suspicious nature, e.g., to sell products of adult nature, and phishing attempts

Features Overview

- **User:** avg. # posts, account age, # subscribed lists, verified account, **posts' interarrival time**, default profile image
- **Text:** # hashtags, # uppercases, # emoticons, # URLs, **sentiment**, avg. word embedding score, hate and curse scores
- **Network:** **popularity** (# follower, # friends), **reciprocity**, avg. **power difference with mentioned users**, **hubs and authority**, influence (**eigenvector centrality**, closeness centrality), communities (**clustering coefficient**, louvain modularity)

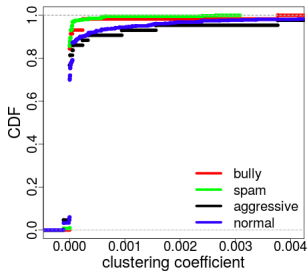
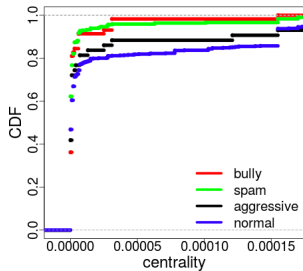
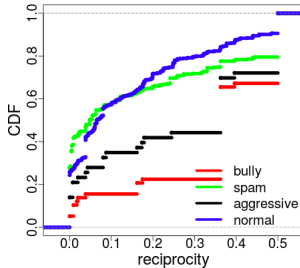
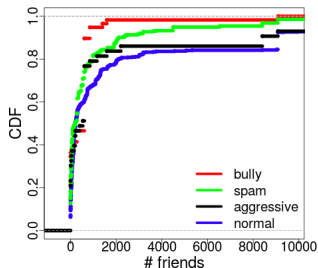
User and Text Features



User and Text Features - Findings

- Normal users sign up to more lists than the other types of users
- Aggressors and bullies post more URLs
- Aggressors and bullies have a propensity to use more hashtags within their tweets
- Clear distinction among the sentiment of aggressors and the other user classes
- Bullies and aggressors are on Twitter for a long time

Network Features



Network Features - Findings

- Bullies have fewer friends and followers than the other user categories, with normal users having the most friends
- Bully and aggressive users are more similar (i.e., higher number of reciprocities) than the normal or spam users
- Influence: bullies follow similar behavior with spammers / aggressors are more similar to normal users
- Bully users, similarly to the spam ones, are less prone to create clusters in relation to aggressive and normal users

Experimentation Phases

- **Detecting offensive classes:** 4-classes classification, i.e., bully, aggressive, spam, and normal users
- **Classifying after spam removal:** 3-classes classification, i.e., bully, aggressive, and normal users



Results

	Prec.	Rec.	ROC
bully	0.411	0.432	0.893
(STD)	0.027	0.042	0.009
aggressive	0.295	0.118	0.793
(STD)	0.054	0.078	0.036
spammer	0.686	0.561	0.808
(STD)	0.008	0.010	0.002
normal	0.782	0.883	0.831
(STD.)	0.004	0.005	0.003
overall (avg.)	0.718	0.733	0.815
(STD)	0.005	0.004	0.031

	Prec.	Rec.	ROC
bully	0.555	0.609	0.912
(STD)	0.018	0.029	0.009
aggressive	0.304	0.114	0.812
(STD)	0.039	0.012	0.015
normal	0.951	0.976	0.911
(STD)	0.018	0.029	0.009
overall (avg.)	0.899	0.917	0.907
(STD)	0.016	0.019	0.005

- **Detecting offensive classes:** prec. - 71.6%, rec. - 73.32%, acc. - 73.45%, RMSE - 0.3086
- **Classifying after spam removal:** prec. - 89.9%, rec. - 91.7%, acc. - 91.08%, RMSE - 0.2117%
- Most contributing features: user- and network-based

Data Balancing

	Prec.	Rec.	ROC
bully	1	0.667	0.833
aggressive	0.5	0.4	0.757
normal	0.931	0.971	0.82
overall (avg.)	0.909	0.913	0.817

- Random Forest suffers from appropriately handling extremely imbalanced training dataset
- **Over-sampling** (SMOTE) and **under-sampling** (resample)
- Random split of data: 90% for training and 10% for testing
- Accuracy - 91.25%, RMSE - 14.23%

Twitter Reaction to Aggression

	active	deleted	suspended
bully	67.24%	32.76%	0.00%
aggressive	65.12%	20.93%	13.95%
normal	86.53%	5.72%	7.75%

Table: Status check on Nov 2016

	active	deleted	suspended
bully	62.07%	37.93%	0.00%
aggressive	55.81%	25.58%	18.60%
normal	85.01%	6.86%	8.13%

Table: Status check on Feb 2017



Findings

- Bullies and aggressors attack in short bursts - not enough duration or content
- They have a long activity on Twitter
- Bullies are less popular and do not participate in many communities
- Aggressors are more difficult to get identified from classifiers



Questions



This work has been funded by the European Commission as part of the ENCASE project (H2020-MSCA-RISE), under GA number 691025.